Supplemental Materials for "ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information"

Rodney LaLonde, Dong Zhang, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida
lalonde@knights.ucf.edu, dromstons@gmail.com, shah@crcv.ucf.edu

We show more intermediate results in this supplementary material to give the reader a better understanding of our method. In Section 1 we provide results for our method when adjusting the scoring distance for considering a proposed detection as a true positive. In Section 2, we show how ClusterNet and FoveaNet work together to improve the performance. In Section 3, we show qualitative results and receiver-operator curves (ROC) for each AOI on the dataset. Due to the space limit and the large size of the figures, it is not possible to show all the results in the main paper.

1. Choice of Scoring Threshold

To be consistent with current literature, we chose to score our proposed detections as true positives if they fell within 20 pixels of a ground truth annotation (one detection allowed per ground truth, additional detections which do not fall within another annotation's 20 pixel radius are considered false positives). Given the average vehicle size is roughly 9×18 pixels, this means detections could be a fair distance from landing on the pixels associated with a vehicle and still be marked as correct. To address this concern, we re-evaluated all of our experimental results at scoring distances ranging from the 20 pixel threshold all the way to within a single pixel to be considered as a true positive. Figure 1 shows our resulting F1 scores at each scoring distance threshold. As one can see, the F_1 score remains high for all scoring threshold values above $1/2$ the average vehicle length.

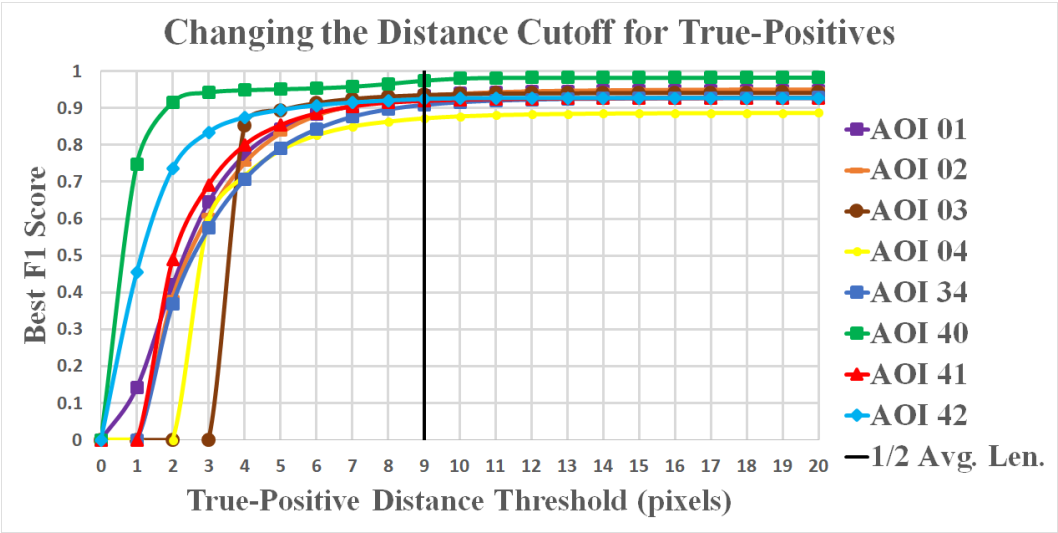


Figure 1: F_1 score as a function of true-positive cutoff distance. The vertical line represents $\frac{1}{2}$ the average vehicle length.

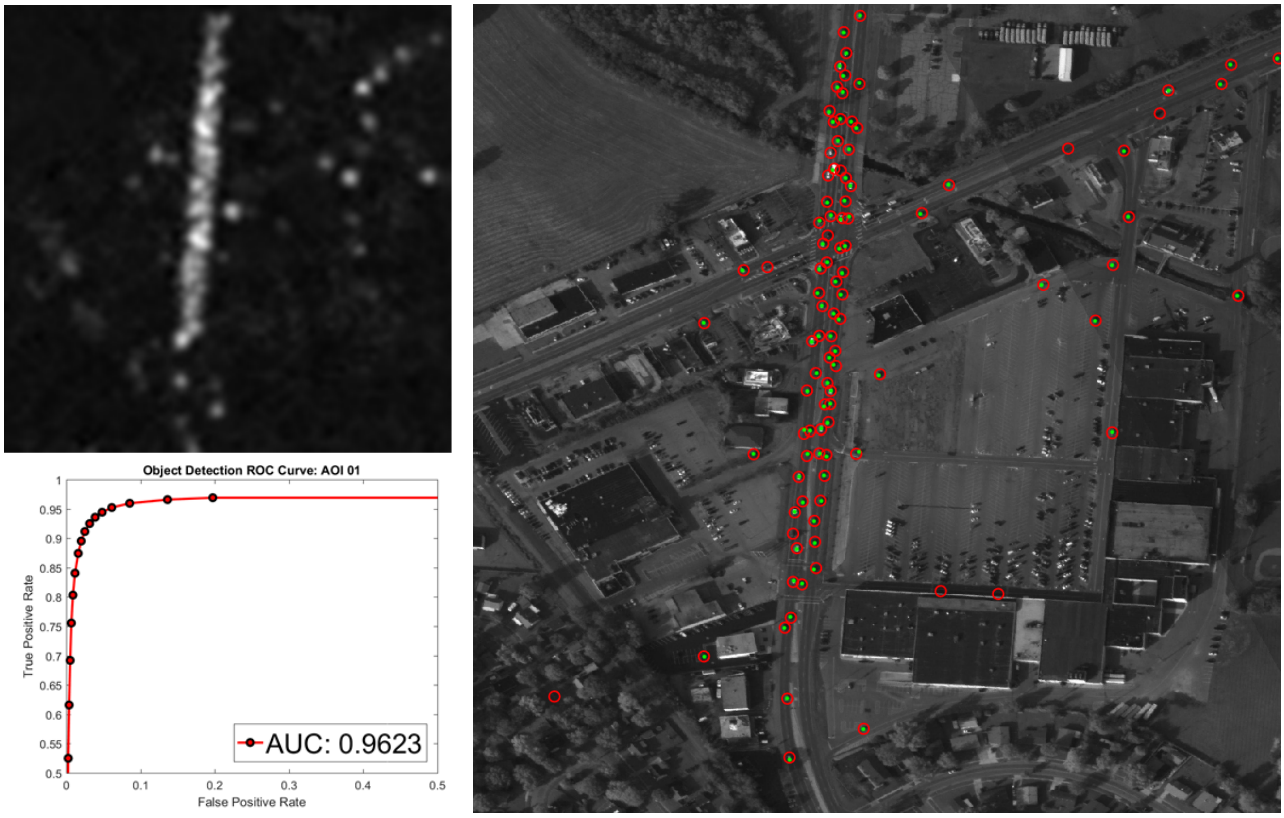
2. Two-Stage CNN Visualized With Qualitative Results

An overview and the performance of different components of the proposed method are shown in Figure 2 (shown below Figure 3a). ClusterNet takes as input a set of video frames, containing a very large search space due to the large size of

each frame. Each high-scoring 4×4 region of ClusterNet's output has the associated region of the input space selected, which is selected based on the propagated receptive field of those 16 output neurons. All low-scoring regions are ignored (set to zero in the output). Working with several to hundreds of megapixel video frames, the frames must be downsampled dramatically early in the network in order to fit within video RAM for deep learning. As a result, localizing individual objects becomes a significant challenge. Each neuron in the output layer of ClusterNet can see anywhere from a single object to none to over 300, depending on object density. This is best illustrated by the magenta boxed region (corresponding to ROOBI 3), where, even in a sparse area of interest (AOI), a single proposed region of objects of interest (ROOBI) contains two objects separated by a significant distance in the original input space. ROOBIs obtained by ClusterNet are then sent through FoveaNet to simultaneously obtain the final locations of all objects of interest in that region to a high degree of accuracy. The example shown obtains final object locations with perfect precision and recall while needing to check only the 9 highest-scoring, of the possible 324, ROOBIs of the output space, saving significant computational time.

3. Qualitative Results for Each AOI and ROC Curves

Figure 3: **Left Column Top:** Output of ClusterNet for the given frame shown at right. **Left Column Bottom:** Receiver operator curves (ROC) to compliment the precision-recall curves in the main paper. True Positive Rate is measured by $TP/(TP + FN)$ where TP is the number of true positive detections and FN is the number of false negative detections. False Positive Rate is measured by FP/NV where FP is the number of false positive detections and NV is the number of vehicles in the ground truth. **Right Column:** Final output of the proposed two-stage framework for example frames for AOIs of the WPAFB 2009 dataset. Red Circles are centered on ground truth coordinates. Green dots are the final predicted object locations by the proposed framework.



(a) AOI 01 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.

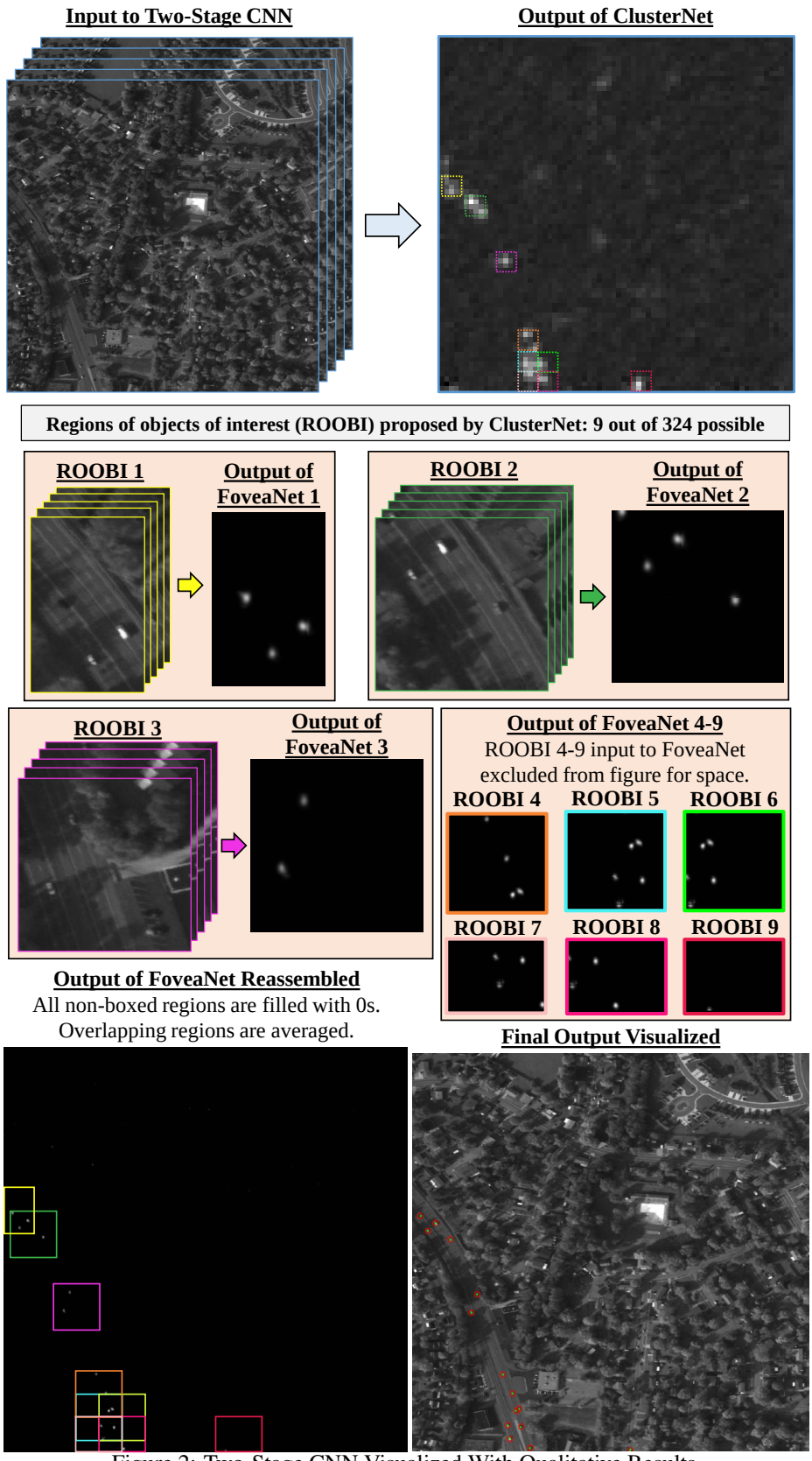
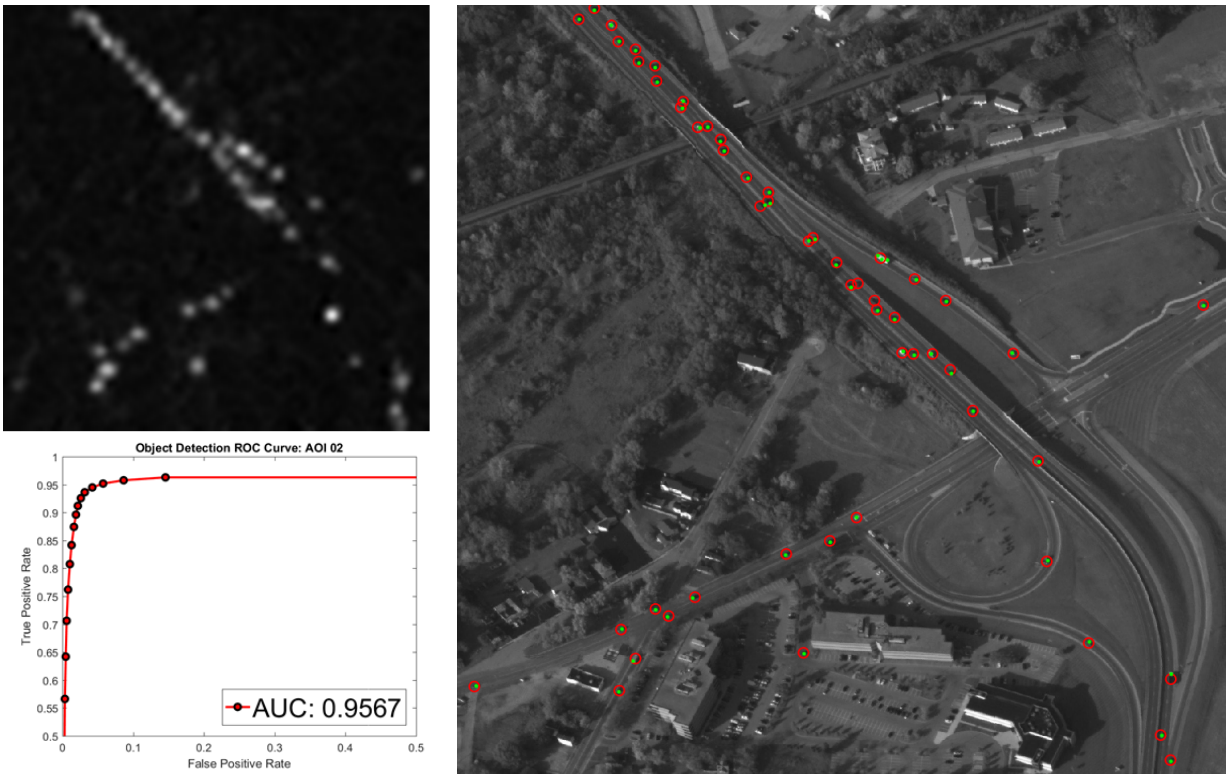
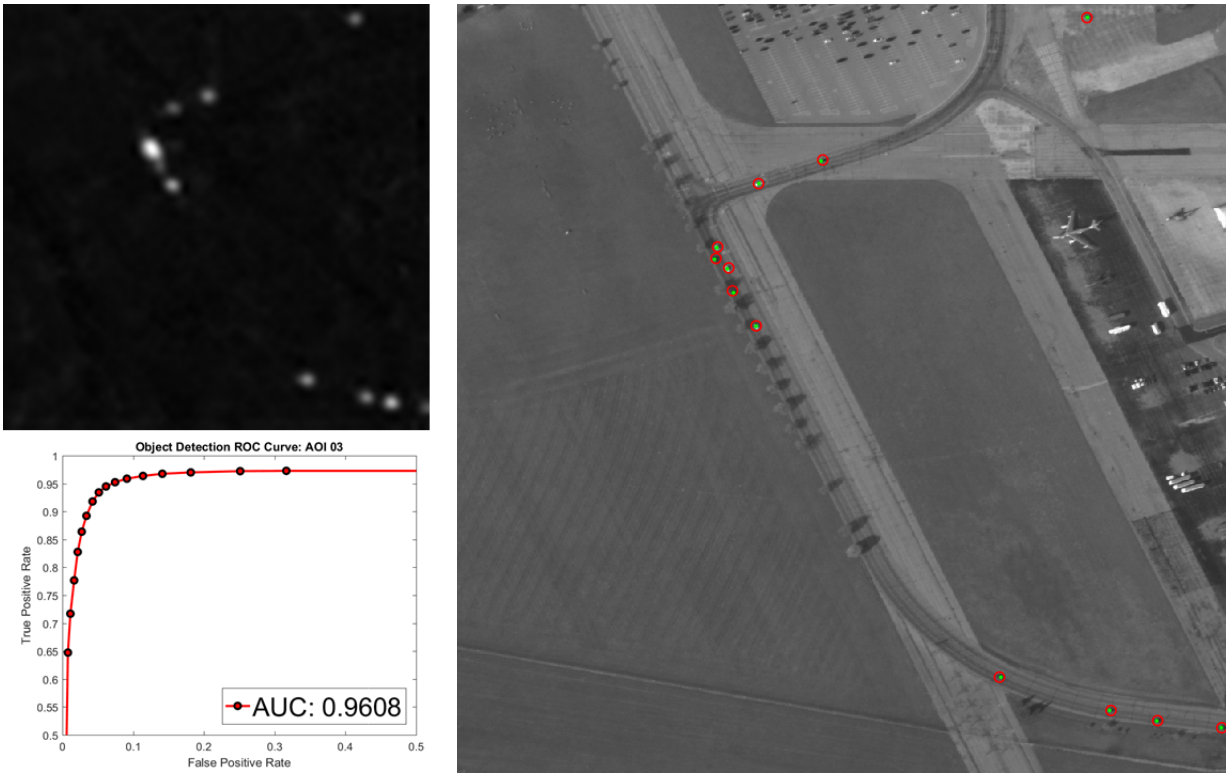


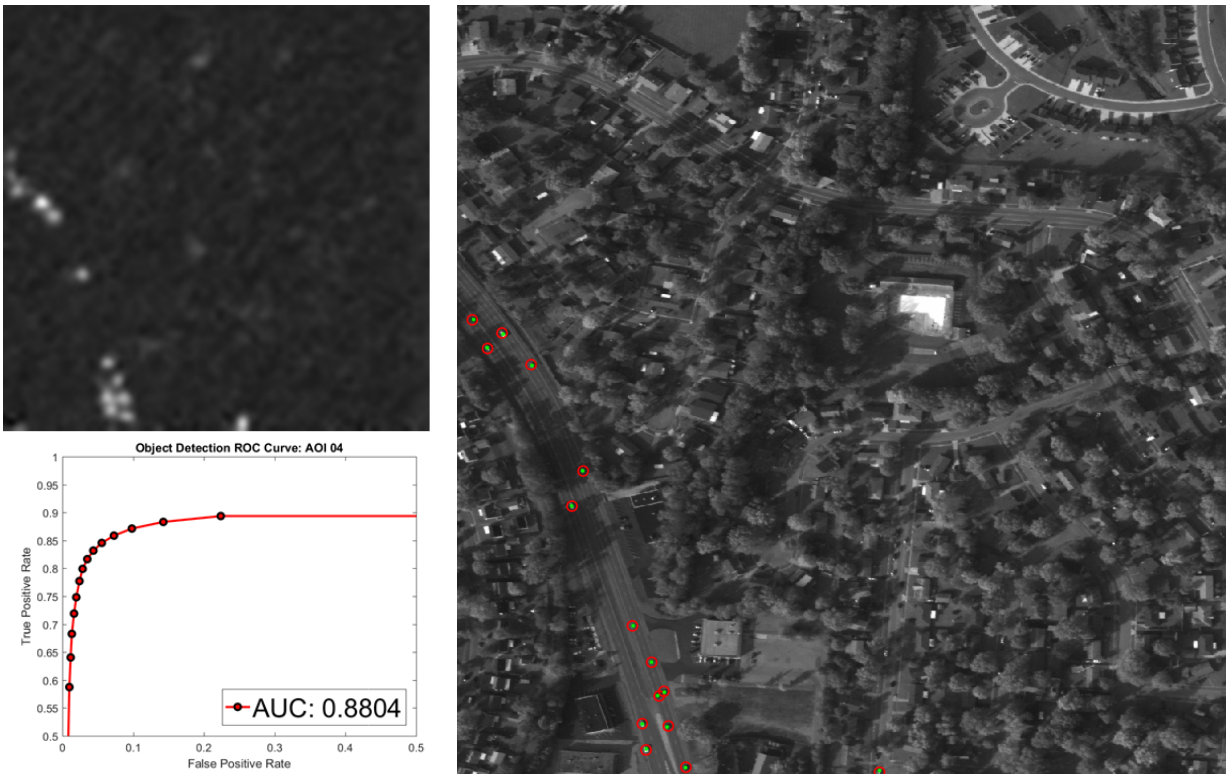
Figure 2: Two-Stage CNN Visualized With Qualitative Results



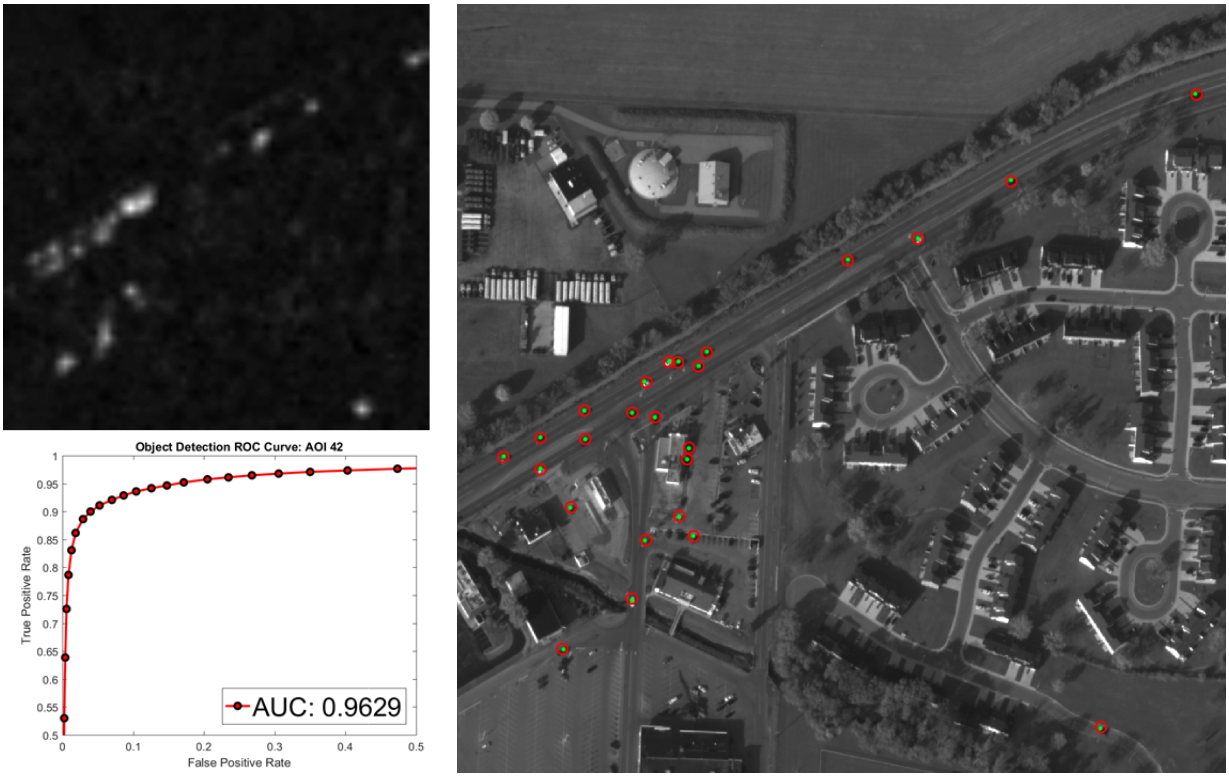
(b) AOI 02 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



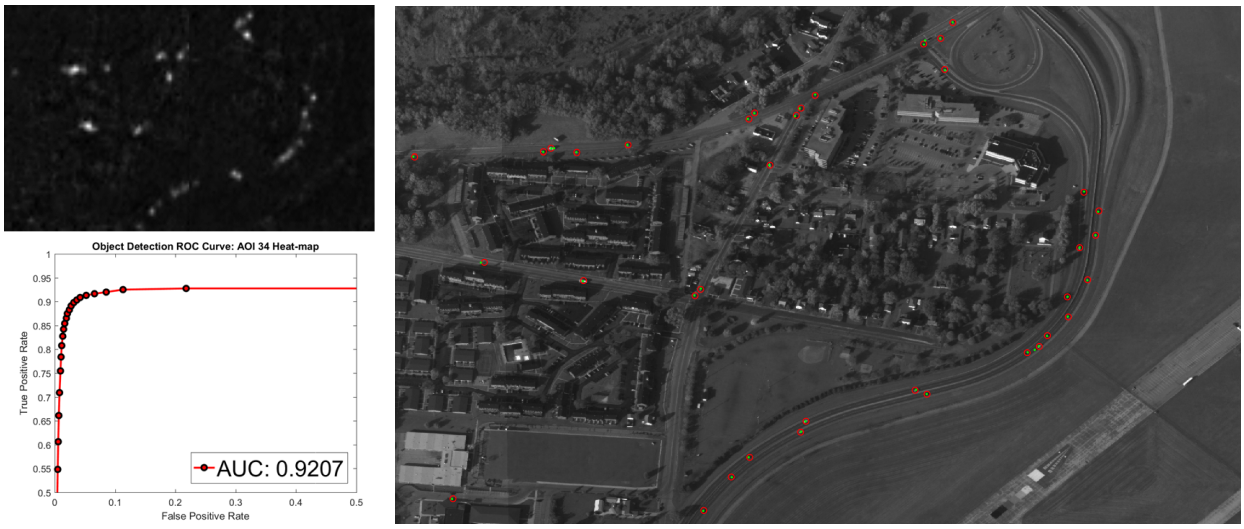
(c) AOI 03 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



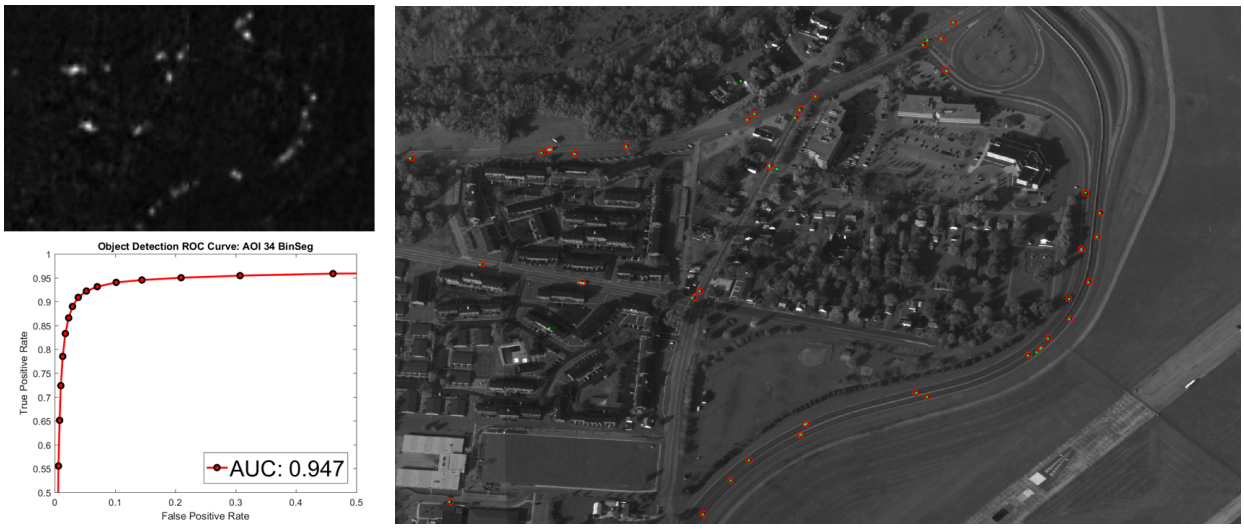
(d) AOI 04 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



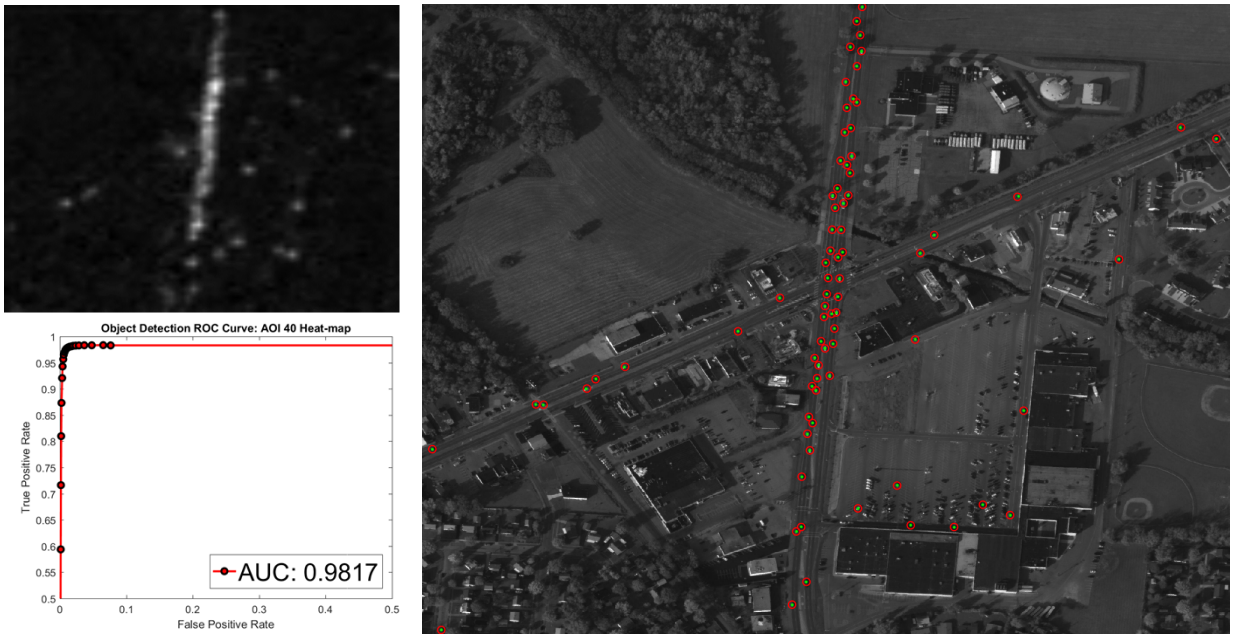
(e) AOI 42 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right. Note AOI 42 contains all ground-truth coordinates, stopped vehicles are not removed.



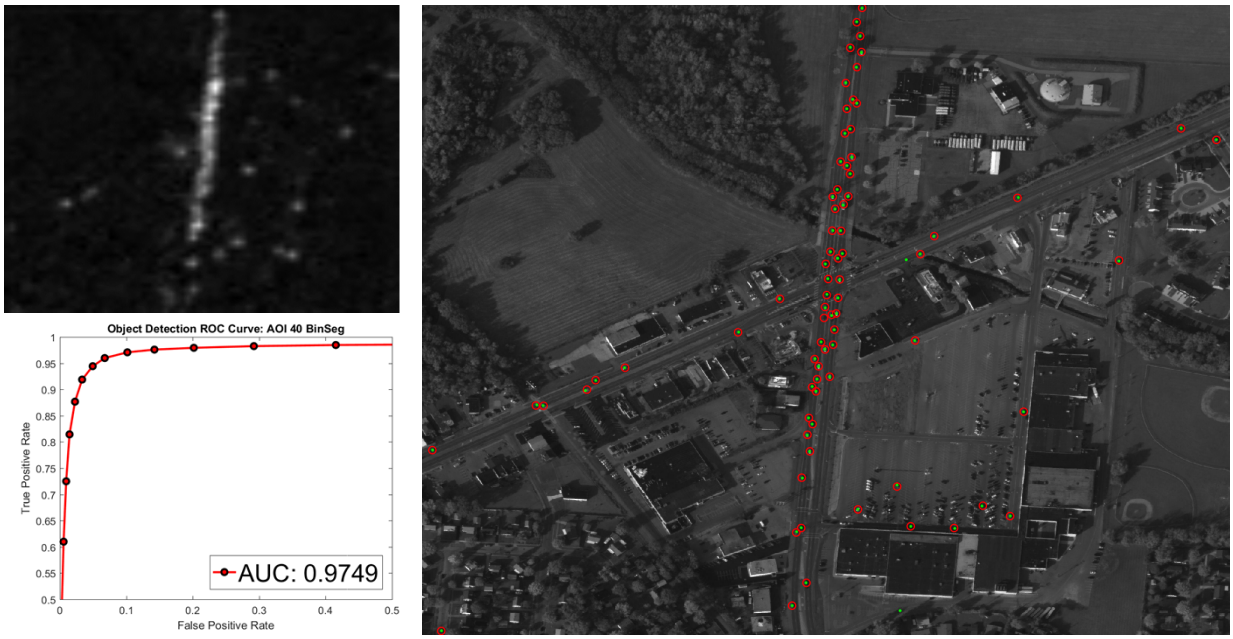
(f) AOI 34 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



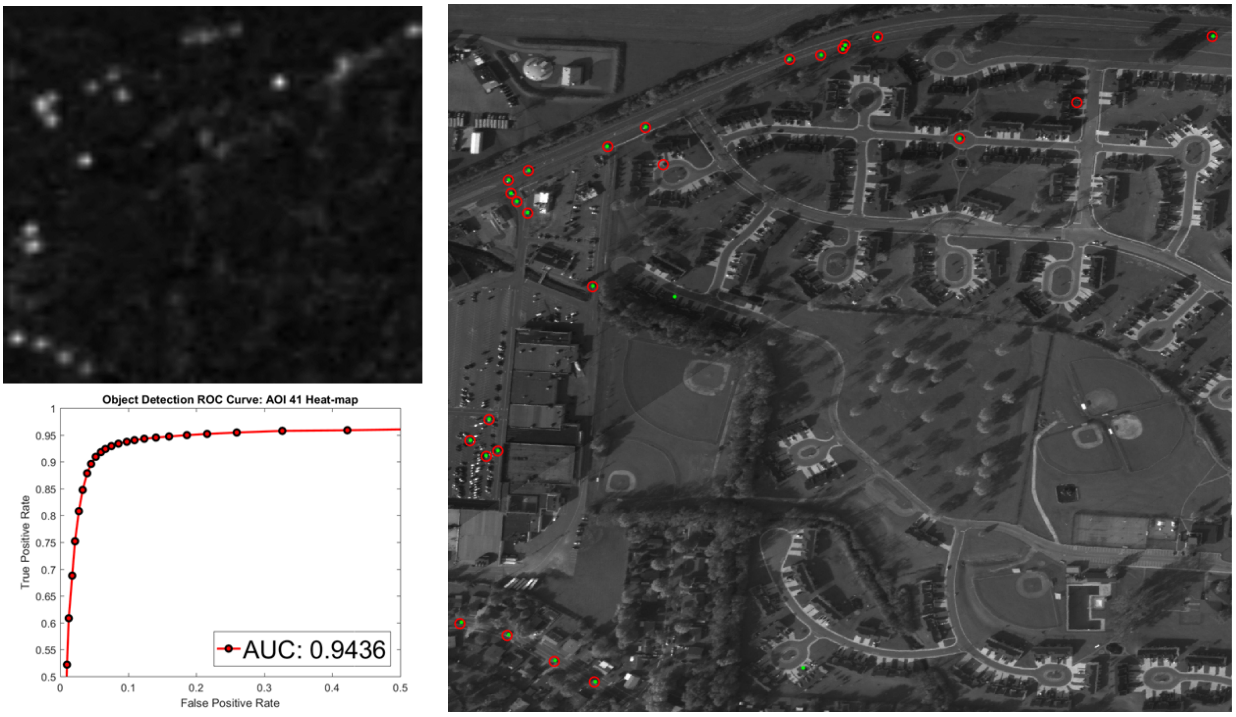
(g) AOI 34 results using 5-frames and the binary segmentation formulation for FoveaNet. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



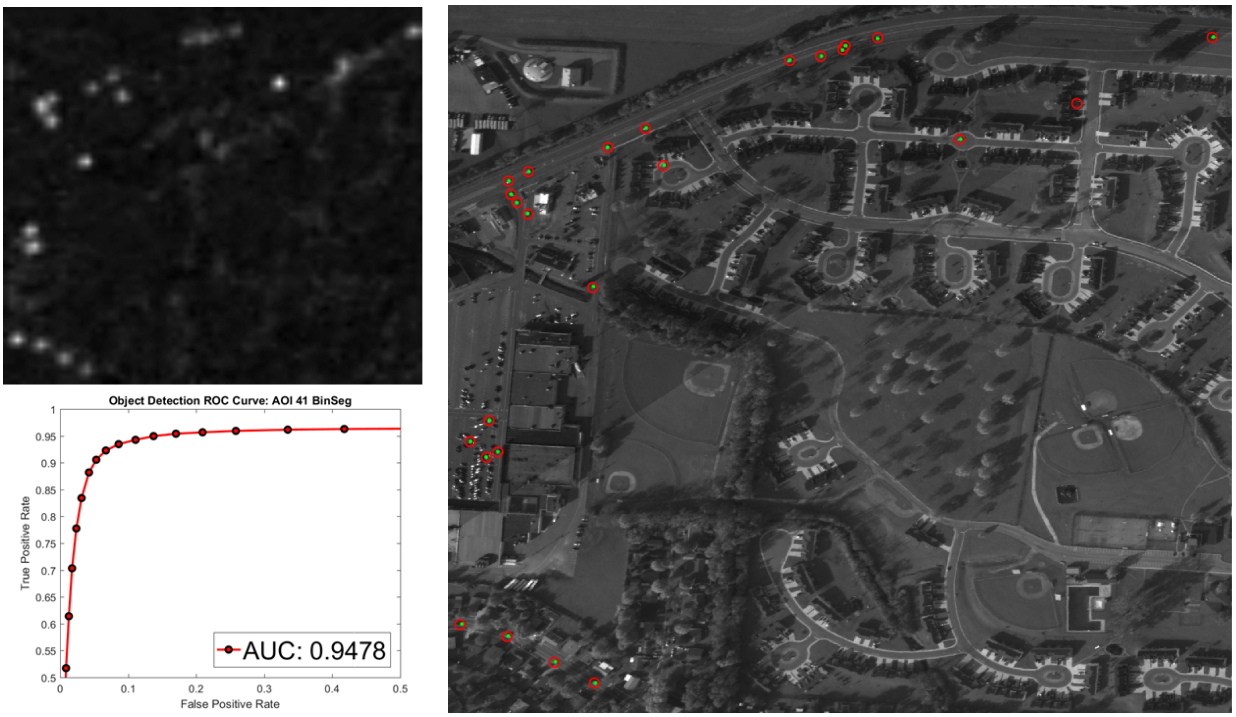
(h) AOI 40 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



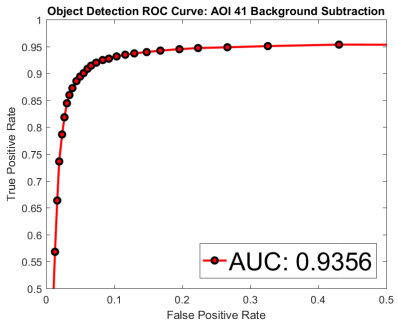
(i) AOI 40 results using 5-frames and the binary segmentation formulation for FoveaNet. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.

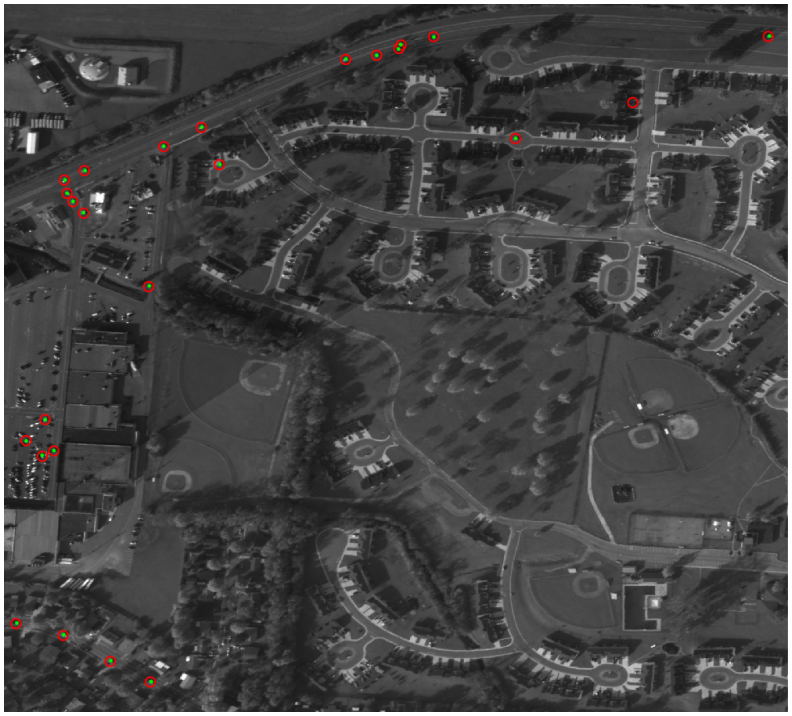


(j) AOI 41 results using 5-frames and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.

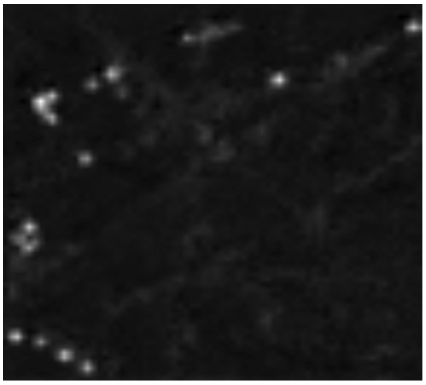


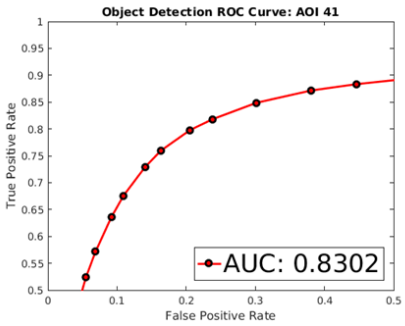
(k) AOI 41 results using 5-frames and the binary segmentation formulation for FoveaNet. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.

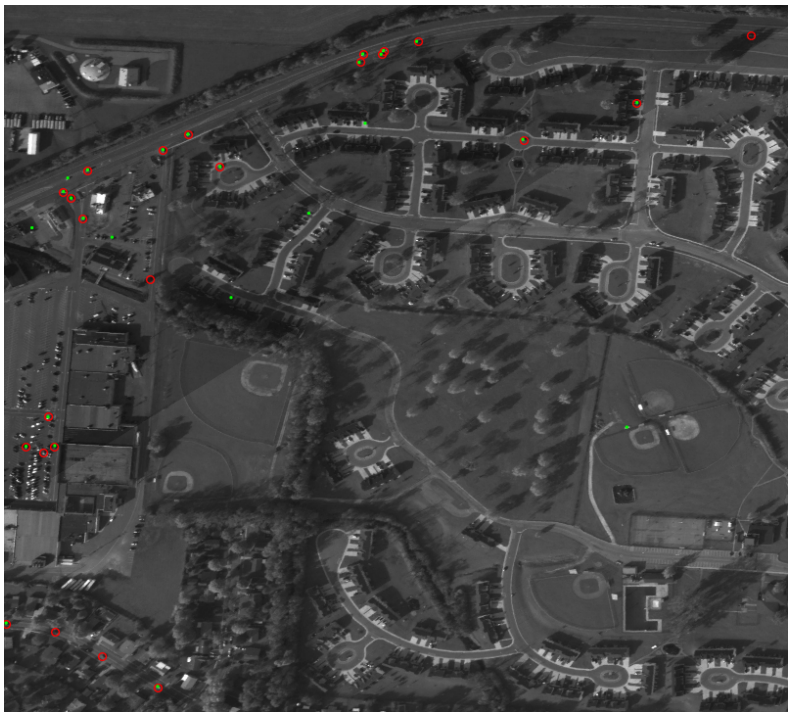




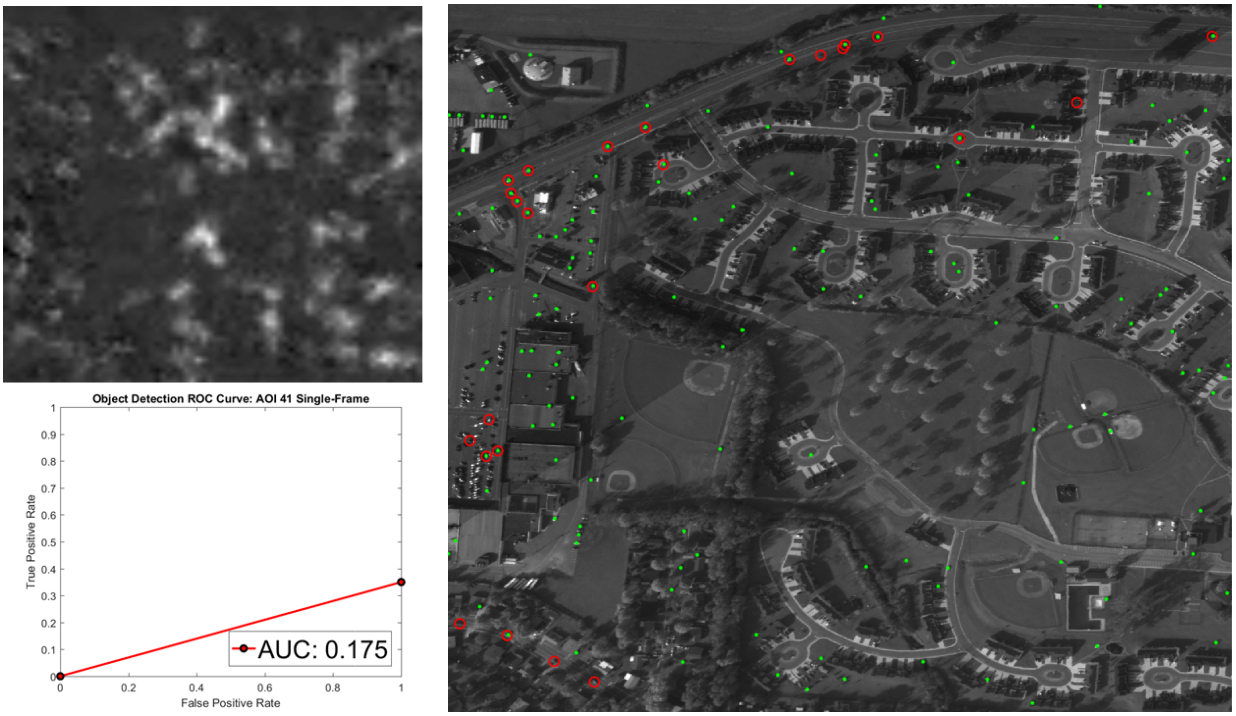
(l) AOI 41 results using the deep learning background subtraction approach.







(m) AOI 41 results using 3-frame and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.



(n) AOI 41 results using 1-frame and the Gaussian heatmap formulation. ClusterNet output shown at left; FoveaNet output and ground-truth shown at right.

(o) AOI 41 results using Faster R-CNN.